PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 20 November 15, 1934 Number 11

THE ACCURACY OF LEAST SQUARES SOLUTIONS

By T. E. Sterne

Harvard College Observatory, Cambridge, Mass.

Communicated September 28, 1934

1. *The General Equations of Least Squares.*—Often we know the results of n observations, $c_1' \ldots c_n'$, such that c_i' may be supposed to have been drawn at random from a normal universe of unknown standard deviation σ (the same for all the observations), and with its mean at $a_i\xi_a + b_i\xi_b + d_i\xi_d + \ldots + m_i\xi_m$. Here the coefficients $a_i \ldots m_i$ are supposed to be known, while the m quantities $\xi_a \ldots \xi_m$ are unknown. We suppose that $n > m$, and the question to be answered is (A): If we know only the values of c_i' and of a_i, b_i, etc., what are the values of the ξ's?

This is the classical problem of the solution of n linear equations of condition in m unknowns, when each of the equations

$$a_i\xi_a + b_i\xi_b + \ldots + m_i\xi_m = c_i', \quad (i = 1, 2, \ldots n) \tag{1}$$

is in error by an amount ϵ_i that is supposed to be drawn at random from a normal population of errors, of constant but unknown standard deviation σ. In practical problems, we may have reason for believing that the errors of different equations do not correspond to the same σ; in this case the usual procedure is to multiply each equation through by the square root of a certain "weight," designed to reduce all the errors ϵ to a common basis σ. We suppose that this has already been done, if necessary, so that the equations (1) to be solved satisfy the conditions specified above. The classical answer to the problem is well known; it is supplied by the method of least squares, and was worked out in all its essentials by Legendre,[1] Gauss[2] and Laplace.[3] That answer is as follows: Find the values x_a, x_b', $\ldots$ x_m', of ξ_a, ξ_b, $\ldots$ ξ_m, respectively, that make $\Sigma_i\delta_i^2$ a minimum. The quantity

$$\delta_i = a_i\xi_a + b_i\xi_b + \ldots + m_i\xi_m - c_i'.$$

The values (called the "least squares values") x_a', x_b', $\ldots$ x_m' are obtained by solving the set of m equations

$$\begin{bmatrix} \Sigma_i a_i(a_i x_a' + b_i x_b' + \ldots + m_i x_m' - c_i') = 0 \\ \vdots \\ \Sigma_i m_i(a_i x_a' + b_i x_b' + \ldots + m_i x_m' - c_i') = 0 \end{bmatrix} \tag{2}$$

or what is the same thing

$$\begin{bmatrix} [aa]x_a' + [ab]x_b' + \ldots + [am]x_m' = [ac'] \\ \vdots \\ [ma]x_a' + [mb]x_b' + \ldots + [mm]x_m' = [mc'] \end{bmatrix} \tag{2'}$$

in which $[aa] = \Sigma a_i^2$, etc.; $[ab] = [ba] = \Sigma a_i b_i$, etc.; and $[ac'] = \Sigma a_i c_i'$, etc. Let $\delta_i' = a_i x_a' + \ldots + m_i x_m' - c_i'$, and define a quantity $s'^2 = (1/n - m)\Sigma\delta_i'^2$. Then the chance that any unknown, ξ_a say, differs from x_a' by an amount lying between $ts'\left|\Delta_{aa}/\Delta\right|^{1/2}$ and $(t + dt)s'\left|\Delta_{aa}/\Delta\right|^{1/2}$ is

$$p(t)dt = \frac{1}{(2\pi)^{1/2}} e^{-\frac{t^2}{2}} dt.$$

Here Δ_{aa} is the complementary minor of x_a' in the determinant Δ whose elements are the coefficients of the unknowns x' in the equations (2').

This is the classical answer to the question (A), correct for very large values of $n - m$. It is not correct for small values of $n - m$, for when $m = 1$ the problem reduces to that of "Student," discussed later. The classical theory has set σ equal exactly to s', whereas σ is really unknown; and we shall presently see that for $n - m$ small, the classical theory overestimates in consequence the accuracy of the "least squares" values. Proceeding rigorously from the suppositions of the first paragraph, and without recourse either to inverse probabilities or to any additional hypotheses, we shall find the exact answer to question (A).

The probability that c_1' will lie between c_1 and $c_1 + dc_1$, c_2' between c_2 and $c_2 + dc_2$, etc., is

$$p(c_1, c_2, \ldots c_n)dc_1 \ldots dc_n = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\Sigma\delta_i^2}{2\sigma^2}} dc_1 \ldots dc_n. \tag{3}$$

From the definition of the δ_i's, given above,

$$\Sigma_i\delta_i^2 = \Sigma_i[a_i(\xi_a - x_a') + b_i(\xi_b - x_b') + \ldots + m_i(\xi_m - x_m') + a_i x_a' + b_i x_b' + \ldots + m_i x_m' - c_i]^2$$

and hence from the definition of the δ_i''s and x''s, and of s', all given above,

$$\Sigma_i\delta_i'^2 = (n - m)s'^2 + \Sigma^2$$

where

$$\Sigma^2 = \Sigma_i[a_i(\xi_a - x_a') + b_i(\xi_b - x_b') + \ldots + m_i(\xi_m - x_m')]^2.$$

We wish to obtain the chance, $p(s, a, \ldots m)dsda \ldots dm$, that s' lies between s and $s + ds$, while the error, $a' = x_a' - \xi_a$, of x_a' lies between a

and $a + da$; the error b' of x_b' between b and $b + db$; etc. We need an expression for the volume element, in the n-dimensional space of the c's, corresponding to $dsdadb \ldots dm$. Consider the points $P(c_1', c_2', \ldots c_n')$ and $P'[(a_1x_a' + \ldots + m_1x_m'), (a_2x_a' + \ldots), \ldots (a_nx_a' + \ldots)]$. Consider also the m lines

$$\begin{aligned}
&1. \quad (c_1/a_1) = (c_2/a_2) = \ldots = (c_n/a_n)\\
&2. \quad (c_1/b_1) = (c_2/b_2) = \ldots = (c_n/b_n)\\
&\vdots\\
&m. \quad (c_1/m_1) = (c_2/m_2) = \ldots = (c_n/m_n).
\end{aligned}$$

Then (1) the line from P to P' has the direction ratios $(c_1' - a_1x_a' - \ldots - m_1x_m'):(c_2' - a_2x_a' - \ldots)$: etc., and is accordingly perpendicular, by equations (2) (the definition of the x''s) to all the lines 1, 2, $\ldots m$. (2) There exists a set of m planes, each normal to one of the lines $1 \ldots m$, and containing the point P':

$$\begin{cases} \Sigma a_i(c_i - a_ix_a' - \ldots - m_ix_m') = 0 \\ \vdots \\ \Sigma m_i(c_i - a_ix_a' - \ldots - m_ix_m') = 0. \end{cases} \tag{5}$$

The line PP' lies in all of these planes and therefore in their intersection, (c). (c) is a plane in $n - m + 1$ dimensions, and is itself of $n - m$ dimensions. It contains the point P', and is normal to all the lines $1 \ldots m$. (3) If $P''(c_1'', c_2'', \ldots c_n'')$ is any point in (c), and if $x_a'', x_b'', \ldots x_m''$ are defined by the equations (of least squares)

$$\Sigma_i a_i(c_i'' - a_ix_a'' - \ldots - m_ix_m'') = 0 \tag{6}$$

etc., then $x_a'' \equiv x_a'$, $x_b'' \equiv x_b'$, etc. For P'' lies in (c) and hence $c_1'', c_2'', \ldots c_n''$ must satisfy all the relations (5). The coefficients of the equations (5) and (6) are identically equal, and hence $x_a'' \equiv x_a', x_b'' \equiv x_b'$, etc. (4) Any point $P'''(c_1''', c_2''', \ldots c_n''')$ whose coördinates lead to the least square values x_a', x_b', etc., lies in (c), since the c''''s satisfy the equations (5). Accordingly, the surface (c) contains all the points whose coördinates lead to the least squares values x_a', x_b', etc., and contains such points only. (5) The (distance)^2 of P from P' is $\Sigma_i(c_i - a_ix_a' - \ldots - m_ix_m')^2 = (n - m)s'^2$. Hence, corresponding to any set of values $s', x_a', x_b', \ldots x_m'$, the point $c_1', c_2', \ldots c_n'$ must lie on a certain sphere S in $n - m$ dimensions, of radius $(n - m)^{1/2}s'$ and center P', and lying wholly in (c). (6) The displacement of P' corresponding to da has direction ratios $a_1:a_2:a_3$, etc.; hence it is parallel to the line 1, and hence it is normal to (c); similarly for the other displacements db, etc. Accordingly, the volume element in the space of the c''s, corresponding to $dsda \ldots dm$ is

$$const \times s^{n-m-1}dsda \ldots dm.$$

The numerical factor depends only upon $n - m - 1$ and upon the angles between the lines $1 \ldots m$, and since $n - m - 1$ and the a_i's, b_i's, etc., are all kept constant, the factor is constant.[4]

The distribution of samples of s', x_a', x_b', etc. (i.e., of s', a', b', etc.), is therefore

$$p(s, a, b, \ldots m)dsda\ldots dm = const' \times s^{n-m-1}e^{-\frac{(n-m)s^2 + \Sigma^2}{2\sigma^2}} dsda\ldots dm \quad (7)$$

where[5]

$$\begin{aligned}\Sigma^2 &= \Sigma_i(a_i a + b_i b + \ldots + m_i m)^2 \\ &= [aa]a^2 + [bb]b^2 + \ldots + 2[ab]ab + 2[ad]ad + \ldots\end{aligned}$$

in which the square bracket expressions have already been defined. If we integrate (7) over all values of $b, \ldots m$, we obtain the distribution of samples of s' and a'

$$p(s, a)dsda = const'' \times s^{n-m-1}e^{-\frac{(n-m)s^2 + a^2|\Delta/\Delta_{aa}|}{2\sigma^2}} dsda \quad (8)$$

where Δ and Δ_{aa} have already been defined. The form of the coefficient of a^2 in the exponent arises from the well-known properties of an essentially positive quadratic form like Σ^2, which is composed only of square terms. If, contrary to hypothesis, σ were known, then (8) would be the equation at which to stop.

Make a Jacobian transformation from variables a,s to t,s where

$$t = \frac{a}{s}\left|\frac{\Delta}{\Delta_{aa}}\right|^{1/2}.$$

Then

$$p(t, s)dtds = const.''' \times s^{n-m}e^{-\frac{(n-m)s^2[1 + t^2/(n-m)]}{2\sigma^2}} dt\,ds$$

is the chance that t will lie between t and $t + dt$, while s' lies between s and $s + ds$. Integrate over all s; then

$$p(t)dt = D\frac{dt}{[1 + t^2/(n-m)]^{\frac{n-m+1}{2}}}, \quad (9)$$

where

$$D = \frac{1}{(n-m)^{1/2}}\frac{n-m-1}{n-m-2}\frac{n-m-3}{n-m-4}\cdots\begin{bmatrix}\cdots\frac{3}{2}\times\frac{1}{2} & \text{(when } n-m \text{ is even)} \\ \cdots\frac{2}{1}\times\frac{1}{\pi} & \text{(when } n-m \text{ is odd)}\end{bmatrix}$$

and does not involve σ.

Equation (9) is the final answer to our problem, and $p(t)dt$ is the chance that the error $x_a' - \xi_a$ of x_a' lies between $ts'\left|\Delta_{aa}/\Delta\right|^{1/2}$ and $(t + dt)s'\left|\Delta_{aa}/\Delta\right|^{1/2}$. The factor $s'\left|\Delta_{aa}/\Delta\right|^{1/2}$ is the classical expression for the "mean" or standard error of x_a', and the results for x_b', etc., can be obtained by merely interchanging the rôles of the symbols. From equation (9), all the properties of the distribution follow. The function defined by (9) differs from the Gaussian distribution in predicting, particularly for small values of $n - m$, a greater probability of large errors than the latter. As $n - m$ is allowed to increase, while t is kept constant, the quantity $p(t)dt$ approaches the Gaussian value of the classical solution, asymptotically, correctly in the term of highest order in $n - m$. It will be seen from (9) that the function $p(t)$ is symmetrical about $t = 0$; that whenever $n \geqslant m + 1$, the "least squares" values given by (2) or (2′) are the most probable values of the unknowns, and when $n \geqslant m + 2$, are also the mean values of the unknowns. For $n = m + 1$, no mean exists. The true root mean square error of a "least squares" value x_a', say, is $s'[(n - m)/(n - m - 2)]^{1/2}\left|\Delta_{aa}/\Delta\right|^{1/2}$ by (9) when $n > m + 2$, and is infinite when $n \leqslant m + 2$. This peculiarity arises from the fact that a fortuitously close concordance of the observations, while improbable, is still sufficiently probable when $n \leqslant m + 2$ to cause $\int_0^\infty t^2 p(t)dt$ to diverge.

The quantity

$$P = \int_{-\infty}^{t} p(t')dt'$$

is the probability of obtaining a result t algebraically smaller than the value of t observed, and corresponds therefore to the "normal error function" of classical theory. P depends upon t and $n - m$, merely. When there is but one unknown (the case of the weighted mean of n observations) this function becomes

$$P = D\int_{-\infty}^{t} \frac{dt}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}$$

of which values have been tabulated by "Student" (in *Metron*, **5**, No. 3, 105 (1925)), in connection with the accuracy of a mean value. We have therefore shown that if these same tables are but entered with arguments t and a value of "Student" 's n' equal to our $n - m + 1$, they will yield directly the value of P in the most general case of least squares, when there are n equations of condition in m unknowns. It appears at once, from an inspection of these tables, that the chance of an error algebraically greater than three mean or standard errors in the value of some unknown.

determined by least squares, is far greater when $n - m$ is small than the chance, 0.0013, predicted by the classical theory. For $n - m$ equal to 24, the actual chance is 0.0031; for $n - m$ equal to 12, the actual chance is 0.0055; for $n - m$ equal to 2, the chance is 0.0478. The factor by which the ordinary mean or standard error should be multiplied to obtain the probable error (i.e., the error as likely as not to be numerically exceeded) has the value unity for $n - m$ equal to 1, and diminishes slowly to the value 0.6745 for $n - m$ infinite.

The development above follows rigorously[6] from only two assumptions: (1) that the errors of the equations of condition are a sample drawn at random from a normal universe of unknown standard deviation σ, and (2) that nothing is known about the values of the several unknowns or of σ except that which may be inferred from the given equations of condition.

I hope that the formulae of this paper may be of some use to those who employ observational material that is liable to accidental errors.

2. *Application to Regression Coefficients.*—The results obtained here are of wide applicability. They include the results of "Student" (*Biometrika*, **6**, 1 (1908); *Metron*, **5**, No. 3, 105 (1925)) for the accuracy of a mean, as the special case that arises when we set m equal to 1 in equation (9). They include, also as a special case, the results obtained by R. A. Fisher (*Journal of the Royal Statistical Society*, **85**, 611 (1922); *Metron*, **5**, No. 3, 102 (1925)) for the accuracy of regression coefficients. We merely take m equal to one more than the number of regression coefficients b in the general equation of multiple regression

$$Y - \bar{y} = b_1x_1 + b_2x_2 + \ldots + b_px_p \qquad (I)$$

where $x_1 \ldots x_p$ are either variates, or functions of one or more variates, measured from their means, in terms of which the regression is expressed (thus the unknowns are here the b 's and $\bar{y}$, the observed quantities being y and the x 's). In this equation $\bar{y}$ is the mean of the dependent variate y, and it is supposed by Fisher that the standard deviation of y, although unknown, is constant for all possible sets of values $x_1 \ldots x_p$. It will be seen that the equations of multiple regression (I) differ from the still more general equations (1) in that in the former, there is always one unknown $\bar{y}$ whose coefficient is unity, while in the latter, all the coefficients are quite unrestricted in value. When all the variates in a regression equation are measured from their means, the term $\bar{y}$ disappears; but there are still really $p + 1$ unknowns. The appearance of the quantity $(n - p)/2$ instead of $(n - p + 1)/2$, as the exponent of the denominator in the distribution function obtained by Fisher, follows therefore as a natural consequence of equation (9) when $m = p + 1$. We have called our problem the problem of "least squares," and have used the usual least

squares notation. In terms of the general equation of multiple regression, (I), the problem we have solved is that of the determination of the distribution of the regression coefficients b when it is supposed that the standard deviation of the dependent variate y is neither constant nor known for different sets $x_1 \ldots x_p$, but that the ratios of such standard deviations, for different sets, are known. When such ratios are all equal to unity, our problem simplifies into Fisher's—a problem which, while general enough for many valuable biological applications, has not been sufficiently general for application to the most frequently arising problems in practical physics and astronomy. The latter lead to systems of equations of condition, like (1), in which there is usually no unknown whose coefficient (even in the original, unweighted, equations) is the same in all the equations; they do not in general lead to systems of equations of the form (I).

[1] Legendre, *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*, p. 72, Paris (1806).

[2] Gauss, *Theoria Motus Corporum Coelestium*, S. 177, Hamburg (1809); *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, Werke, **4**, p. 1, Göttingen (1873).

[3] Laplace, *Théorie Anal. des Prob.*, **2**, Chap. 4 (1812).

[4] If one doubts the validity of geometrical reasoning, it is easy to recast the above argument into a purely analytical form.

[5] It is seen that s' is not correlated with any of the x''s, but that there is normal correlation between the x''s.

[6] After these results were obtained, they were compared with those of Jeffreys (*Proc. Roy. Soc.*, **A138,** 48 (1932)) and found to be the same. Jeffreys, however, obtained his results merely by an arbitrary choice of the distribution $p(\sigma)d\sigma \propto d\sigma/\sigma$ for σ, regardless of the sample; and the defective nature of that procedure was shown by Fisher (*Proc. Roy. Soc.*, **A139,** 343 (1933)) and Bartlett (*Proc. Roy. Soc.*, **A141,** 518 (1933)), all of whose objections are met by the treatment given here.

IDEALS IN ALGEBRAIC RINGS

By Garrett Birkhoff*

Harvard University

Communicated October 15, 1934

1. *Reduction by Ordinary Primes.*—This paper is a study of factorization in algebraic rings, and outlines a very direct attack which goes deep into the problem. Among other things, it shows that the divisibility (or "ideal") structure of an algebraic ring is either entirely canonical or very irregular. It further gives an arithmetic criterion for determining which algebraic integers give rise to canonical ideal structure.

For brevity, detailed proofs are omitted; of course they have been constructed. For brevity also, the nomenclature and notation of footnote 1 will be used without explanation.

Now let x be any algebraic integer, and let

$$f(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n = 0 \qquad (1)$$

be the irreducible equation for x. The rational integers and x generate a ring R, and we shall denote by $L(R)$ the system whose elements are the non-zero ideals $A, B, C, \ldots$ of R, and whose operations are g.c.f., l.c.m., product and ideal quotient. R and $L(R)$ are thoroughly familiar concepts to algebraists, and will be fixed throughout the paper.

By the "index" of an ideal $A \epsilon L(R)$ we mean the order of the residue ring R/A. Any A must contain a rational integer $a \neq 0$, and so must have a finite index dividing a^n. To every prime p consequently corresponds the system $L_p(R)$ of ideas whose index is a power of p.

Moreover it is true, although the proof is not easy, that $L(R)$ is essentially** the direct product of the $L_p(R)$. We can summarize this symbolically by writing

$$L(R) = L_2(R) \times L_3(R) \times L_5(R) \times \ldots \qquad (2)$$

and confine our attention to the various individual $L_p(R)$.

2. *Strong Isomorphism and Strong Semi-Isomorphism.*—Let $I \supset J$ and $K \supset L$ be any four ideals of R. A (1, 1) correspondence between the elements of the residue rings I/J and K/L will be called a "strong isomorphism" if and only if it preserves addition and multiplication, and is preserved under multiplication by an arbitrary $z \epsilon R$; it will be called a strong semi-isomorphism if and only if it preserves addition and is preserved under multiplication by an arbitrary $z \epsilon R$. In the first case we write $I/J \cong K/L$; in the second, $I/J \simeq K/L$.

THEOREM 1: If $I, J < I$ and $K < I$ are distinct ideals satisfying $(J, K) = I$, and if $I/J \simeq I/K$, then L exists, an ideal satisfying $(L, J) = (L, K) = I$ and $[L, J] = [L, K] = [J, K]$.

We form L from those residue classes $[J, K] + j + k$ of $[J, K]$ for which $j \epsilon J$ and $k \epsilon K$ correspond under the given strong semi-isomorphism.

3. *Irregular Case.*—The situation concerning the ideal of R/pR is totally clarified by the remark that we can apply Euclid's algorithm to polynomials modulo p, so that the ideals of R/pR correspond (1, 1) with the divisors of $f(x)$ modulo p which generate them. We now prove

THEOREM 2: Suppose that $g^2(x)$ divides $f(x)$ modulo p, that $g(x)$ is irreducible modulo p, and yet that the ideal $g(x)R \not\supset pR\ (p^2)$. Then the ideals of R do not satisfy

$$[(J, K), (J, L)] = (J, [K, L]). \qquad (3)$$

Set $I = (g(x)R, pR)$, $J = (g(x)R, p^2R)$, and $K = (g^2(x)R\ pR)$. We obtain ideals satisfying the hypotheses of Theorem 1, and therefore satisfying $[(J, K), (J, L)] = (J, K) < J = (J, [K, L])$, contrary to (3).

4. *Canonical Case.*—It can be proved that if $g^2(x)$ does not divide $f(x)$ modulo p while $g(x)$ is an irreducible factor of $f(x)$ modulo p, then $(g(x)R, pR)$ is a principal ideal generated by some $g(x) + kp$ (p). But if $g(x)R \supset pR(p^2)$, then $(g(x)R, pR)$ is a principal ideal, modulo any p^n. Finally, it is true that every largest ideal of $L_p(R)$ contains pR. Consequently, applying paragraph one of §3, we see that unless the hypotheses of Theorem 2 are satisfied for some $g(x)$, every largest ideal of $L_p(R)$ is a principal ideal, modulo any p^n.

But it can be proved that if T is a principal ideal in $L_p(R)$, then the transformation $I \longrightarrow TI$ carries $L_p(R)$ into the subsystem of ideals contained in T, isomorphically relative to g.c.f. and l.c.m. Using induction on index, we obtain

THEOREM 3: If no $g(x)$ satisfies the hypotheses of Theorem 2, then $L_p(R)$ is "canonical"—that is, isomorphic with the system of the products of a set of ordinary primes, relative to g.c.f., l.c.m., product and quotient.

5. *Consequences.*—Now we use our relation (2) to pass from $L_p(R)$ to $L(R)$. But we first note that any canonical system of ideals is "regular" in the sense of Grobner[3]—that is, that if $A \subset B$, then $A : (A : B) = B$—and that by Theorem 8 of footnote 2 any "regular" system of ideals satisfies (3). Therefore

THEOREM 4: The properties of satisfying (3), of being regular and of being canonical, are effectively equivalent in R, and all are equivalent to the property that if $g(x)$ is any irreducible factor of $f(x)$ modulo p, then either $g^2(x)$ does not divide $f(x)$ modulo p, or else $(g(x)R, p^2R) \supset pR$.

The author intends to publish elsewhere complete proofs of the assertions made above, together with related theorems and applications to the arithmetics of specific algebras. As an example of the results proved, we shall cite

THEOREM 5: The number of different simple factors J/K in $L_p(R)$ relative to strong isomorphism, is twice the number of irreducible factors of $f(x)$ modulo p.

* Society of Fellows, Harvard University.

** Technically, the "subdirect product" in the sense of footnote 2, §3.

[1] O. Ore, "Abstract Ideal Theory," *Bull. Am. Math. Soc.*, **39**, 728–45 (1933).

[2] G. Birkhoff, "On the Lattice Theory of Ideals," *Ibid.*, **40**, 613–19 (1934).

[3] W. Grobner, "Uber irreduziblen Ideale in kommutativen Ringe," *Math. Ann.*, **110**, 161–94 (1934).

V_3 IN R_6 OF $[1,2]_a$

BY NATHAN KAPLAN

DEPARTMENT OF MATHEMATICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Communicated October 5, 1934

I. *Introduction.*—For an introduction to our methods, we refer the reader to the paper "V_3 in R_6."[1]

II. *V_3 in R_6 of* [1,2].—Following the classification given in the above paper, these V_3 in R_6 are characterized by:

$$V_{\underset{pq}{a}} = 0 \tag{1.1}$$

$$h_{\underset{4}{ab}} = hi_a i_b \tag{1.2}$$

$$h_{\underset{5}{ab}} = hi_a i_b \tag{1.3}$$

$$h_{\underset{6}{ab}} = hi_a i_b + hi_a i_b. \tag{1.4}$$

Here i_1, i_2, i_3, i_7 are arbitrary but i_3 is perpendicular to i_7. The equations of Gauss, Codazzi and Ricci are

$$K_{abcd} = \Sigma_p (h_{ac} h_{bd} - h_{ad} h_{bc}) \tag{2.1}$$

$$h_{a[b,c]} = \Sigma_q V_{[b} h_{c]a} \tag{2.2}$$

$$V_{[a,b]} = h.^{c}_{[a} h_{b]c} - \Sigma_r V_{[a} V_{b]}. \tag{2.3}$$

Using (I), from the Ricci Relations (2.3), we find that V_3 of [1,2] contains (4) subcases according as

$$i_1 = i_2;\ i_1 \text{ is perpendicular to } i_3, i_7 \tag{3.1}$$

$$i_1 = i_2;\ i_1, i_3, i_7 \text{ lie in the } \textit{same } R_2 \tag{3.2}$$

$$i_1 \text{ is } \perp \text{ to } i_2;\ i_1, i_2, i_3, i_7 \text{ lie in the } \textit{same } R_2 \tag{3.3}$$

$$i_1 \text{ is } \perp \text{ to } i_2;\ i_1, i_3, i_7 \text{ lie in the } \textit{same } R_2;\ i_2 \text{ is } \perp \text{ to } R_2. \tag{3.4}$$

We denote these classes by $[1,2]_a$; $[1,2]_b$; etc., according as to whether (1), (3.1); (1), (3.2); etc., are satisfied. In this paper, we shall study $[1,2]_a$.

III. *V_3 in R_6 of* $[1,2]_a$.—The V_3 in R_6 of this class possess the following quantities:

$$V_{\underset{pq}{a}} = 0 \tag{1.1}$$

$$h_{\underset{4}{ab}} = h i_{\underset{4}{a}} i_{\underset{1}{b}} \tag{1.2}$$

$$h_{\underset{5}{ab}} = h i_{\underset{5}{a}} i_{\underset{1}{b}} \tag{1.3}$$

$$h_{\underset{6}{ab}} = h i_{\underset{6}{a}} i_{\underset{2}{b}} + h i_{\underset{7}{a}} i_{\underset{3}{b}} \tag{1.4}$$

$$i_{\underset{1}{}}, i_{\underset{2}{}}, i_{\underset{3}{}} \text{ are mutually orthogonal.} \tag{1.5}$$

For normal curvature properties, we find

(2.1) the normal curvature vectors lie in an R_2 through each point P.

(2.2) the asymptotic directions have no special significance.

(2.3) to each direction $i_{\underset{a}{}}$ in the R_2 perpendicular to $i_{\underset{1}{}}$, there corresponds a plane of conjugate directions.

From the Codazzi Relations, we find

$$i_{\underset{1}{}}, i_{\underset{2}{}}, i_{\underset{3}{}} \text{ are triply orthogonal } V_2 \text{ building} \tag{3.1}$$

$$V_2 \text{ of } (i_{\underset{2}{}}, i_{\underset{3}{}}) \text{ is geodesic in } V_3 \tag{3.2}$$

$$i_{\underset{1}{}} \text{ is a geodesic of } V_3 \tag{3.3}$$

$$\text{using a coördinate system along } i_{\underset{1}{}}, i_{\underset{2}{}}, i_{\underset{3}{}} \tag{3.4}$$

$$h_{\underset{4}{}} = h_{\underset{4}{}}(u'), \; h_{\underset{5}{}} = h_{\underset{5}{}}(u')$$

$$h_{\underset{6}{}} = h_{\underset{6}{}}(u^2, u^3), \; h_{\underset{7}{}} = h_{\underset{7}{}}(u^2, u^3) \tag{3.5}$$

$$\frac{\partial h_{\underset{6}{}}}{\partial u^3} = \frac{h_{\underset{7}{}} - h_{\underset{6}{}}}{h_2} \frac{\partial h_2}{\partial u^3}; \quad \frac{\partial h_{\underset{7}{}}}{\partial u^2} = \frac{h_{\underset{6}{}} - h_{\underset{7}{}}}{h_3} \frac{\partial h_3}{\partial u^2} \tag{3.6}$$

$$ds^2 = (du')^2 + (h_2 du^2)^2 + (h_3 du^3)^2 \tag{3.7}$$

$$h_2 = h_2(u^2, u^3); \; h_3 = h_3(u^2, u^3).$$

For the Gauss Relation, Ricci Tensor and K of V_3, we find from (1.1)

$$\frac{\partial}{\partial u^2}\left(\frac{1}{h_2}\frac{\partial h_3}{\partial u^2}\right) + \frac{\partial}{\partial u^3}\left(\frac{1}{h_3}\frac{\partial h_2}{\partial u^3}\right) = -h_{\underset{6}{}} h_{\underset{7}{}} (h_2 h_3)^2 \tag{4.1}$$

$$K_{ad} = -h_{\underset{6}{}} h_{\underset{7}{}} (i_{\underset{2}{a}} i_{\underset{2}{d}} + i_{\underset{3}{a}} i_{\underset{3}{d}}) \tag{4.2}$$

$$K = -2 h_{\underset{6}{}} h_{\underset{7}{}}. \tag{4.3}$$

Since the ∞' V_2 of $(i_{\underset{2}{}}, i_{\underset{3}{}})$ are geodesic in V_3, denoting the fundamental quantities of these ∞' V_2 by $(^-)$, we find from the Gauss Relations for V_2 in V_3

$$\overline{K}_{2332} = K_{2332} \tag{5.1}$$

$$\overline{K}_{ab} = K_{ab} \ (a, b = 2, 3) \tag{5.2}$$

$$\overline{K} = K. \tag{5.3}$$

Hence (4.1) is the Gauss Relation for the ∞' V_2 $\underset{2}{(i,} \ \underset{3}{i)}$ in V_3. From a simple analysis of the Weingarten Formulae, denoting any V_2 of $\underset{2}{(i,} \ \underset{3}{i)}$ by x^λ (u^2, u^3, c) and V_3 by y^λ (u^1, u^2, u^3), we find

$$\left(\frac{\partial^2 y^\lambda}{\partial u^a \partial u^b}\right)_{u^1 = c} = \frac{\partial^2 x^\lambda}{\partial u^a \partial u^b} \ (a, b = 2, 3) \tag{6.1}$$

$$\underset{p}{h_{ab}})_{u^1 = c} = \underset{p}{\overline{h}_{ab}} \ (p = 4, 5, 6) \tag{6.2}$$

$$\underset{1}{i^\lambda} = \frac{\partial y^\lambda}{\partial u^1} \tag{6.3}$$

$$\Gamma'_{ab} = \underset{1}{\overline{h}_{ab}} = 0. \tag{6.4}$$

Hence (4.1) is also the Gauss Relation for V_2 in R_6. Finally, a simple analysis shows that the ∞' V_2 of $\underset{2}{(i,} \ \underset{3}{i)}$ have the following fundamental quantities.

$\bar{g}_{ab}$ of a particular V_2 are arbitrary; the ∞' V_2 are applicable (7.1)

$$\underset{p}{\overline{h}_{ab}} = 0 \quad (p = 1, 4, 5) \ (a, b = 2, 3) \tag{7.2}$$

$$\underset{6}{\overline{h}_{ab}} = \underset{6\,2\,2}{hi_a i_b} + \underset{7\,3\,3}{hi_a i_b} \tag{7.3}$$

$$\underset{pq}{\overline{V}_a} = 0 \quad (p, q = 4, 5, 6, 1). \tag{7.4}$$

Furthermore a simple calculation shows that (3.6) are the Codazzi Relations of these ∞' V_2 in R_6. Hence, the ∞' V_2 in R_6 are completely specified by the above.

From (7.2), (7.3) by the use of a very simple theorem, we find that the ∞' V_2 lie in ∞' R_3.

Hence the V_3 *of* $[1,2]_a$ *are generated by* ∞' *identical* V_2 *lying in* ∞' R_3. *The direction of generation* $\underset{1}{i}$ *is orthogonal to* ∞' R_3.

From the above geometry, one sees that the integral equations of these V_3 are of the type

$$y^\lambda = a^\lambda_b(u^1) \times^b (u^2, u^3) + w^\lambda \ (u^1) \tag{8.1}$$

$$(\lambda = 1, 2, \ldots 6) \quad (b = 1, 2, 3)$$

[1] *Proc. Nat. Acad. Sci.*, **20**, 489–494 (1934).

THE FUNDAMENTAL GROUP OF A GROUP MANIFOLD

BY P. A. SMITH

MATHEMATICS DEPARTMENT, BARNARD COLLEGE, COLUMBIA UNIVERSITY

Communicated October 10, 1934

It is well known that the fundamental group G of an r-parameter continuous group $\mathfrak{G}$ is abelian.[1] We shall prove (in outline) that *G cannot contain more than r linearly independent elements.*[2]

Suppose on the contrary that it is possible to choose $r + 1$ linearly independent elements $\omega_1, \ldots, \omega_{r+1}$; this assumption will eventually lead to a contradiction. First let us form the r-parameter continuous covering group $\mathfrak{G}^0$ belonging to the subgroup G^0[3] generated by the ω_i's. The fundamental group of $\mathfrak{G}^0$ is isomorphic[4] to G^0; hence it has a set of generators $\omega_1^0, \ldots, \omega_{r+1}^0$ which we may take to be independent paths beginning and ending at e, the identity element of $\mathfrak{G}^0$.

Now let $(\theta_1, \ldots, \theta_r)$ be an arbitrary point of an r-dimensional torus $\mathfrak{G}^*$ and let ω_i^* $(i = 1, \ldots, r)$ be the closed path in $\mathfrak{G}^*$ obtained by allowing θ_i to increase from 0 to 2π and holding the remaining θ's equal to 0. We may suppose that the first r of the ω_i^0's are defined by single-valued continuous functions, say $\tau_1(\theta_1), \ldots, \tau_r(\theta_r)$ where $\tau_i(0) = \tau_i(2\pi) = e$. Then the function

$$f(\theta_1, \ldots, \theta_r) = \tau_1(\theta_1)\tau_2(\theta_2) \ldots \tau_r(\theta_r)$$

(where the product on the right is the group product in $\mathfrak{G}^0$) defines a single-valued continuous mapping of $\mathfrak{G}^*$ on $\mathfrak{G}^0$. Moreover f transforms ω_i^* into ω_i^0 $(i = 1, \ldots, r)$ and thus transforms G^*—the fundamental group of $\mathfrak{G}^*$—isomorphically into the subgroup of G^0 generated by $\omega_i^0, \ldots, \omega_r^0$. This subgroup is of infinite index and hence by the results of H. Hopf,[5] the absolute grade[6] of f is zero. From this it follows from another theorem of Hopf[7] that f can be continuously modified to a mapping f' such that the image set $f'(\mathfrak{G}^*)$ is nowhere dense in $\mathfrak{G}^0$. Since $\mathfrak{G}^0$ is locally cartesian,[2,3] it follows from dimension theory that $\dim f'(\mathfrak{G}^*) < r$.

Let Ω^0 and Ω^* be the universal covering spaces of $\mathfrak{G}^0$ and $\mathfrak{G}^*$. The mapping f' of $\mathfrak{G}^*$ on $\mathfrak{G}^0$ induces in an obvious manner a mapping F' of Ω^* on Ω^0, and it is not difficult to see that

$$\dim F'(\Omega^*) = \dim f'(\mathfrak{G}^*) < r. \qquad (1)$$

The spaces Ω^0 and Ω^* are of course non-compact. They can, however, be made compact by the addition of sets Δ^0 and Δ^* of ideal elements, defined by methods due to Alexandroff.[8] The remainder of the proof now depends on the following result: the sets Δ^0 and Δ^* can be so chosen that:

(*a*) They will be homeomorphic to r- and $(r-1)$-spheres, respectively, and $\Omega^* + \Delta^*$ will be homeomorphic to a closed r-dimensional spherical region;

(*b*) There will exist an $(r-1)$-sphere Δ^1 in Δ^0 and a homeomorphism T between Δ^1 and Δ^* such that if $\xi_1^*, \xi_2^*, \ldots$ is any sequence of points in Ω^* converging to δ^*, a point of Δ^*, then $F'(\xi_i^*) \longrightarrow T(\delta^*)$.[9]

Suppose that we now extend the domain of definition of F' to the whole of $\Omega^* + \Delta^*$ by defining $F'(\delta^*) = T(\delta^*)$ for every δ^* in Δ^*. Then on account of (*a*), (*b*) and (1), F' defines a single-valued continuous transformation of a closed r-dimensional spherical region into a set of dimension less than r, and in particular, transforms the bounding $(r-1)$-sphere homeomorphically. The existence of such a transformation would be disturbing to the intuition, and indeed its mathematical impossibility follows immediately from a theorem of Alexandroff.[10]

[1] First proved by Schreier, *Abh. math. Seminar.*, Hamburg, **5**, 15–32 (1927).

[2] A detailed proof will appear elsewhere; it is carried out for the case in which $\mathfrak{G}$ is a finite or infinite complex which is a simple r-circuit and which has a denumerable complete set of neighborhoods which are r-cells.

[3] H. Hopf (Zur Topologie, etc., zweite Teil, *Math. Ann.*, 102 (1929)) has defined the covering space belonging to a given subgroup of the fundamental group, and Schreier (loc. cit.) has defined the covering continuous group of a continuous group. A natural combination of these definitions leads to the definition of $\mathfrak{G}^0$. It can readily be shown that $\mathfrak{G}^0$ satisfies the conditions stated in note 2 if $\mathfrak{G}$ does.

[4] Hopf, loc. cit., p. 572.

[5] Loc. cit., p. 584.

[6] Defined by Hopf, loc. cit., p. 582.

[7] Loc. cit., p. 605.

[8] See Menger, Dimenstheorie, pp. 182–189.

[9] The proof that Δ^0 and Δ^* can be defined with dimensions r and $r-1$ depends on the fact that G and G^0 are generated by $r+1$ and r independent generators. The proof that a T can be chosen which transforms Δ^* homeomorphically into a subset of Δ^0 and is related to F' in the way stated, depends on the fact that f, and therefore f', transforms G^* isomorphically into a subgroup of G^0.

[10] "Untersuchungen über Gestalt, etc.," *Ann. Math.*, **30** (1928–29), p. 175, Kor. III.

ON THE FOUNDATIONS OF A CONSTRUCTIVE THEORY OF DISCRETE COMMUTATIVE ALGEBRA

BY H. S. VANDIVER

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated October 5, 1934

During several years past the writer has attempted to build up a theory of commutative algebra involving only constructive proofs. No notion of continuity, the Dedekind cut, Cantor's theory of real numbers or the Zermelo postulate is employed, hence, the word "discrete" in the title of this paper.

Before attempting to write an extended account of this theory I shall describe briefly here some concepts and results along these lines which may have some elements of novelty. In addition to the character of the proofs involved, as indicated above, the theory as it is developing has certain other main features as follows:

I.—The consideration of the arithmetic of the natural numbers in such a way that a certain *finite arithmetic* is introduced at the same time, which is associative and commutative, but in which the *cancellation law of addition does not hold*. These two arithmetics are made the starting points from which all of the commutative algebras considered are built up, using the ideas of polynomials involving a finite number of indeterminates, residue classes and homomorphism.

II.—The development of a theory of *semi-groups* and employment of the results for the investigation of *quasi-groups*, groups and various types of algebra. Here a semi-group will be briefly defined as a system of elements closed under an associative operation. A quasi-group is a semi-group such that if $CoA = CoB$ or $AoD = BoD$ for any A, B, C and D in the set we infer in each case $A = B$, where (o) indicates the operation of the semi-group.

III.—Generalization of the concept of ring to obtain other types of associative algebras. In defining these systems the notion of semi-group is found very convenient.

IV.—Extension of parts of the rich theory of fields to commutative rings, employing concepts which enable us to obtain for rings natural and simple extensions of theorems involving fields. Particular attention to the theory of finite rings, so far rarely considered in the literature. All the algebraic systems employed are studied from the standpoint of the rings and fields contained in them, as well as being considered as parts of systems obtained by adjunction of new elements. In the theory of fields this point of view affords a natural approach to the Galois theory.

V.—Building up a theory of real algebraic fields and the ordering of the

elements of such fields, employing modifications and simplifications of ideas due to Kronecker[1] and Schatunovsky.[2] The real algebraic field is obtained from the consideration of a polynomial $f(x)$ with coefficients in an ordered field F such that there are elements a and b of F such that $f(a)f(b) < 0$, $b > a$. We then associate at least one symbolic root α of $f(x) = 0$ with this interval and say that α generates a real field with $b > \alpha > a$. By carrying out a process which enables us to obtain all possible isolated intervals of the type (a, b) we finally obtain a theorem, which takes the place of the ordinary fundamental theorem of algebra, to the effect that we may, in a finite number of steps, adjoin a finite number of real algebraic fields to F so that in the enlarged field, $f(x)$ decomposes into factors of degree < 3.

1. We now give more details concerning some of the topics mentioned above. The finite arithmetic referred to in I will be considered in detail in another paper[3] and consists of a system of m distinct elements

$$C_1, C_2, \ldots, C_i, \ldots, C_m \tag{1}$$

such that $C_{m+1} = C_i$ and which combine under addition and multiplication, as do the natural numbers, except that the order relations do not all hold, that is, from

$$C_a + C_b = C_a + C_k$$

we cannot always infer that

$$C_b = C_k.$$

How the algebras we consider are built up from this arithmetic as well as the arithmetic of the natural numbers we shall indicate briefly; first, by showing how to adjoin a zero element to each of the two arithmetics just mentioned.

Let α be a fixed natural number or a fixed element of (1); then an expression of the form

$$a_{i_1}\alpha^{i_1} + a_{i_2}\alpha^{i_2} + \ldots + a_{i_n}\alpha^{i_n} \tag{2}$$

where the i's denote distinct natural numbers is said to be a *polynomial* in the *indeterminate* α, of the *first kind*, with *coefficients*

$$a_{i_1}, a_{i_2}, \ldots, a_{i_n}$$

denoting natural numbers if α denotes a natural number, and denoting elements of (1) if α denotes an element of (1). There as yet being no zero element in our system, we must speak of a polynomial of the *second kind* as one which has the character of the first kind except that it contains a term not involving α. The polynomial (2) is said to be formally or identically equal to

$$b_{k_1}\alpha^{k_1} + b^{k_2}\alpha^{k_2} + \ldots + b_{k_n}\alpha^{k_n} \tag{3}$$

if the k's form a permutation of the i's and if $i_s = k_t$, then $a_{i_s} = b_{k_t}$. A similar definition will apply to polynomials of the second kind. We shall employ the symbol ($\equiv$) for formally equal, unless a modulus is mentioned and then it will read "congruent to." We define addition and multiplication of polynomials of either type in the way usually done.

If $g(x)$ and $h(x)$ each denote polynomials in an indeterminate x and if they are related in any one of the three ways following:

$$g(x) + A_1(x)f(x) \equiv h(x) + B_1(x)f(x)$$
$$g(x) \equiv h(x) + B_2(x)f(x)$$
$$g(x) \equiv h(x)$$

then $f(x)$ is said to be congruent to $h(x)$, modulo $f(x)$, or

$$g(x) \equiv h(x) \pmod{f(x)}.$$

We find that if $g(x)$ is of the first kind, then

$$g(x) \equiv x \pmod{x}.$$

If $h(x)$ is of the second kind, then

$$h(x) \equiv b \pmod{x}$$

where b does not involve x. We note the properties $a + x \equiv a \pmod{x}$; $ax \equiv x \pmod{x}$. Here a and b denote natural members or elements of (1) depending on the x and the coefficients of the polynomials involved. Hence, if we classify polynomials in x according to their residues, modulo x, the class corresponding to x has the property of a zero element in the set of classes.

We may now include zero with our natural numbers and adjoin in a somewhat similar way the negative integers, employing the modulus $x + 1$, and adjoin the fraction $1/a$, where a is an integer not zero, by using the modulus $ax - 1$. This method, in the last two cases, was suggested by Kronecker[4] and developed by Wedderburn.[5]

We may illustrate what we shall mean by a constructive proof as follows:

Consider an integer $m > 1$, and divide m by each positive integer less than it. If no exact divisor is found, we say that m is prime. If we find a c such that $m = ck$ with k an integer, then we say that m is composite. Since this *finite algorithm* leads us to one of two conclusions, we say that an integer > 1 is prime or composite and *under these circumstances we shall be free to employ the reductio ad absurdum argument and the law of the excluded middle*, in order to show that some integer is prime. Our aim in the present theory will be to attempt to make all the quantities and concepts employed as definite as the idea of prime integer, as the latter was brought out above.

As is well known, Kronecker emphasized the view-point that if a quan-

tity was to be employed in algebra it should be shown, in a finite number of steps, to be expressible in terms of rational numbers. Here, briefly speaking, we shall regard any quantity as existing if it is obtainable from the natural numbers, or (1) using the idea of residue class, analogous to the adjunction of zero, as above described, or if it can be determined in a finite number of steps in terms of these existing elements. This point of view is closely connected with that of Schatunovsky (l. c.) who considered the foundations of algebra and discussed the law of the excluded middle.

Following the constructive point of view as described, we must, if we discuss reducibility of polynomials, be able to set up in advance a finite algorithm which enables us to decide if a given polynomial is reducible or irreducible. If the coefficients of a polynomial $f(x)$ are rational, we have an algorithm of this type due to Kronecker.[5] If $f(x)$ is irreducible, of degree > 1, then the residue classes, modulo $f(x)$, form a field defined by α, where $f(\alpha) = 0$, and because of this isomorphism we say, as usual, that α *exists*. Hence,

$$f(x) = \phi(x)\,(x - \alpha)$$

where $\phi(x)$ has coefficients in the field defined by α, say $F[\alpha]$. We now consider from the constructive standpoint the determination, in a finite number of adjunctions, of a field in which $f(x)$ decomposes into linear factors, the fundamental *decomposition field* (Zerlegungskörper). To carry this out we must accomplish the equivalent of finding linear factors of $\phi(x)$. But $\phi(x)$ has coefficients not in the rational field. Also, the known method[6] of finding in a finite number of steps the irreducible factors in $F[\alpha]$ of any polynomial $g(x)$ with coefficients in $F[\alpha]$ presupposes, among other things, the decomposition of $f(x)$ into linear factors. To avoid this difficulty the writer has devised a method of determining the irreducible factors of $\phi(x)$ in $F[\alpha]$ based on first indicating a finite number of rings, which after adjunction to $F[\alpha]$ gives a ring R in which $f(x)$ decomposes into linear factors.

2. As to semi-groups, this concept is a natural one in the study of rings since the elements of a ring form a semi-group under multiplication, but not always a group or even a quasi-group. Suppose a semi-group S contains an element E such that $EA = A$ for any A of S, then E is called a *left-identity* element of S and if $AE = A$ for any A in S it is called a *right-identity* element of S. (Here we are writing XY for XoY.) If an element is both a right- and left-identity element of S it is called an *identity* element of S. If there is an element O_l in S such that $O_lA = O_l$ for any A in S, then O_l is said to be a *left-annulator* of S and if $AO_r = O_r$ for any A, then O_r is said to be a *right-annulator* of S. Also, if an element O is both a right- and left-annulator of S, it is called an *annulator* of S. It is easy to show that S cannot have more than one identity or more than one annulator.

If S has an identity element E and if there exist elements A, A' and A'' in S such that $AA' = E$ and $A''A = E$, then A is said to be a unit element of S. If U is a unit and A any element of S, then AU or UA are each said to be associates of A. These last two definitions agree with those usually employed in connection with the multiplicative semi-group of an algebraic field and the multiplicative semi-group of square matrices of order n.

If C is an element of S such that from $CA = CB$ for any elements A and B of S, then C is said to be a *left-cancellable* element of S, and if the order is reversed, a *right-cancellable* element. If C is both right- and left-cancellable, it is called *cancellable*. Other elements are termed *non-cancellable*. The product of any two cancellable elements may be shown to be cancellable, and if the semi-group S is commutative (Abelian), that is, $AB = BA$ for any A and B in S, then the product of any element of S by a non-cancellable element is non-cancellable.

In another paper by the writer[3] a necessary and sufficient condition that a finite semi-group contain a cancellable element will be obtained.

Referring to the fundamental separation of a group G into co-sets with respect to a sub-group G_1, this result may be extended to the case *where G is a quasi-group* and G_1 a group included in G.

A concept concerning semi-groups which results in unifying a number of theorems in algebra and number theory is that of BASIS. Suppose we have an Abelian semi-group S and an element A in S. If S contains an identity element E and A is a unit, then we write $A^{-n} = A_1^n$ where $AA_1 = E$ and $A^0 = E$, and we say that the positive and negative integers and zero are the *exponents appertaining to A*. If in the same type of S we have a non-unit N, then we write $N^0 = E$ and the positive integers and zero appertain to N. Lastly, if S contains no identity element, then we say that the positive integers are the exponents appertaining to any element A.

Suppose that in S there exists a set of elements

$$A_1, A_2, A_3, \ldots$$

such that any element of S may be expressed in the form

$$A_{i_1}^{a_1} A_{i_2}^{a_2} A_{i_3}^{a_3} \ldots A_{i_s}^{a_s}$$

where each a is an exponent appertaining to the corresponding A, and the i's are natural numbers. Assume also that if

$$A_{h_1}^{x_1} A_{h_2}^{x_2} \ldots A_{h_t}^{x_t} = A_{h_1}^{y_1} A_{h_2}^{y_2} \ldots A_{h_t}^{y_t}$$

where the h's are integers and each x and y appertain to the corresponding A's then

$$A_{h_c}^{x_c} = A_{h_c}^{y_c}; \ (c = 1, 2, \ldots, t);$$

and if an A appears on one side only of a similar relation then its exponent

is zero. *Then the A's are said to form a basis of the semi-group S.* If the A's are infinite in number, then the basis is said to be infinite, otherwise, finite.

We immediately see from the above that *the well-known factorization theorems in number-theory are included in theorems concerning basis systems of a semi-group.* For example, the theorem that a rational integer > 1 decomposes into prime factors uniquely may be stated in the form: The rational integers > 1 form, under multiplication, a quasi-group with a unique infinite basis. From the present view-point many factorization theorems in number theory in which the phrase is used, "aside from unit factors" now may be translated into theorems in which the behavior of the unit factors is definitely provided for. In connection with the decomposition of polynomials with coefficients in the rational field F, Gauss's lemma and related results may all be stated conveniently as theorems concerning the possible basis systems in the semi-group formed by multiplication of polynomials of this type.

[1] *Werke*, Bd. 3, halbband 1, pp. 263–273; H. B. Fine, *Bull. Amer. Math. Soc.* (1913).

[2] Algebra as a study of congruences with respect to functional moduli (Russian) Odessa (1913?), Chap. 8. I am indebted to Dr. A. E. Ross for an English translation of this.

[3] Note on a simple type of Algebra in which the cancellation law of addition does not hold. *Bull. Amer. Math. Soc.* (1935).

[4] *Werke*, Bd. 3, halbband 1, pp. 260–262.

[5] *Ann. Math.*, **24**, 263–264 (1923).

[6] Kronecker, *Werke*, Bd. 2, pp. 258–9; Van der Waerden, *Moderne Algebra*, Erster Teil, pp. 130–1 (1930).

---

*FUNCTIONALITY IN COMBINATORY LOGIC**

By H. B. Curry

Department of Mathematics, The Pennsylvania State College

Communicated September 20, 1934

1. *Introduction.*—In an attempt to resolve the foundations of logic and mathematics into their elements, it has occurred to several persons that certain notions, ordinarily taken as primitive, could be analyzed into constituents of much simpler nature. Among such notions are, on the one hand, various processes of substitution, and the use of variables generally; and, on the other hand, the categories of logic—such as proposition, propositional function and the like—together with the intuitions by which we tell what entities belong to them.

For a theory concerned with an analysis of these notions I have proposed the name *combinatory logic* (*Amer. Jour. Math.*, **52**, 511 (1930)). This is

a formal theory based on a primitive frame (i.e., set of primitive ideas, axioms and rules of procedure) of such great simplicity that even comparatively simple inferences, such as are ordinarily made by a substitution process can be decomposed into a large number of the elemental steps represented by the rules of procedure; it furthermore does not postulate any such notions as variable, although all the inferences ordinarily made by the use of variables can—when suitable definitions have been made, of course—be made as compound inferences within the system. The development of this theory, to the point where these statements relative to variables can be proved, is contained in a series of papers culminating in "Apparent Variables from the Standpoint of Combinatory Logic" (*Ann. Math.*, 2nd ser., **34**, 381–404 (1933)). To this paper, and to the earlier ones cited therein, the reader is referred for the details of the theory and for the notation.

The object of the present investigation is to obtain a similar analysis of the function concept. To make a more precise statement I must make the following explanations. We are concerned with statements of the form (using E. H. Moore's terminology) "f is a function on X to Y"; or (for functions of several variables) "f is a function on $X_1X_2 \ldots X_m$ to Y." The following type of question suggests itself: Suppose we have given certain entities $f_1, f_2, \ldots, f_n$ concerning each of which a statement of the above form has been made; suppose further that g is an entity derived from $f_1, \ldots, f_n$ by substitution or other such processes; then, what statement of the above form can we infer concerning g? We make inferences of the above form intuitively; what this paper asserts is that when certain formal constituents are added to the primitive frame of combinatory logic, then these inferences can be made abstractly within the extended system.

This formalization of the function notion is an important step toward the aims of combinatory logic (see *Amer. Jour. Math.*, **52**, 511). In the first place it gives *ipso facto* an analysis of the intuitions by which we classify entities into categories. For if the category of a given combination of entities is determinate, it is because the combination is constructed by substituting in certain functions entities which combine with those functions to give new entities of determinate character. It is only necessary to specify what the fundamental categories shall be, and to which ones the primitive entities belong. Thus the present theory provides an analysis of a class of inferences which are ordinarily made either tacitly or by virtue of complex rules which justify the inferences without giving a true analysis of them. In the second place this investigation has a bearing on the contradictions. For many of these contradictions appear to arise from applying the rules, appropriate to a certain category of entities, to an entity which seems to belong to that category but in reality

does not do so. In such cases when we attempt to formalize the proof that the entity does belong to that category the contradiction dissolves. (See §§ 5 and 6 below.)

All the developments of this paper are independent of Ax. Π_0. (See reference above.)

2. *Axioms for Abstract Functionality.*—Let F denote an entity whose interpretation is such that FXY represents, in the notation of combinatory logic, the category of functions on X to Y, while the formula $\vdash FXYf$ represents the statement that f belongs to that category. This F may be called the *functionality relation*, it will be taken as a new primitive idea.

Let entities F_n be defined as follows:

Definition: $F_0 \equiv I$

$F_1 \equiv F$

$F_{n+1} \equiv [x_1, x_2, \ldots, x_n, y, z]\, F_n x_1 x_2 \ldots x_n\, (Fyz).$

Then F_n has the same interpretation relative to functions of n variables that F has relative to functions of one variable; i.e., $F_n X_1 X_2 \ldots X_n Y$ represents the category of functions on $X_1, X_2, \ldots, X_n$ to Y, while the fact that f belongs to that category is expressed by the formula $\vdash F_n X_1 X_2 \ldots X_n Yf$.

The axioms for F are, then, as follows:

Axiom F. $\vdash (x, y, z)(Fxyz \supset (u)(xu \supset y(zu)))$

Axiom (FB). $\vdash (x, y, z)F(Fxy)(F(Fzx)(Fzy))B$

Axiom (FC). $\vdash (x, y, z)F(F_2yxz)(F_2xyz)C$

Axiom (FW). $\vdash (x, y)F(F_2xxy)(Fxy)W$

Axiom (FK). $\vdash (x, y)Fy(Fxy)K$

Axiom (FP)$_1$. $\vdash (x, x', y)(P_1^* x'x \supset P_1^*(Fxy)(Fx'y))$

Axiom (FP)$_2$. $\vdash (x, y, y')(P_1^* yy' \supset P_1^*(Fxy)(Fxy'))$

Axiom (FΠ). $\vdash (x)F(F\lambda\mathrm{Pr})\mathrm{Pr}(P_1^* x)$

where Pr stands for proposition. The last axiom is of a somewhat more dubious character than the others.

3. *Consequences of These Axioms.*—The following are the principal theorems.

3.1. (Immediate consequence of the definition in §2.)

For $m, n = 0, 1, 2, 3, \ldots$

$$\vdash F_{m+n} = [x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n, z]F_m x_1 x_2 \ldots x_m (F_n y_1 y_2 \ldots y_n).$$

I.e.: the two notions—"function (of $m + n$ variables) on $x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n$ to z," and "function (of m variables) on $x_1, x_2, \ldots, x_m$ to functions (of n variables) on $y_1, y_2, \ldots, y_n$ to z" are identical.

3.2. (General substitution theorem, consequence of Axioms F to (FK), inclusive.) *If* —$(1)\mathfrak{X}$ *is a combination of variables* $t_1, t_2, \ldots, t_p, x_1, x_2, \ldots,$

x_m, ($p \geqq 0$, $m \geqq 0$), and $\alpha_1, \alpha_2, \ldots, \alpha_m, \delta$ are combinations of constants and the variables $t_1, t_2, \ldots, t_p$, such that

$$\vdash (t_1, t_2, \ldots, t_p) F_m \alpha_1 \alpha_2 \ldots \alpha_m \delta([x_1, x_2, \ldots, x_m]\mathfrak{X}),$$

(2) *$\zeta_1, \zeta_2, \ldots, \zeta_m$ are combinations of constants and the variables $t_1, t_2, \ldots, t_p, y_1, y_2, \ldots, y_n$, and $\beta_1, \beta_2, \ldots, \beta_n$ are combinations of constants and the variables $t_1, t_2, \ldots, t_p$, such that, for $i = 1, 2, \ldots, m$,*

$$\vdash (t_1, t_2, \ldots, t_p) F_n \beta_1 \beta_2 \ldots \beta_n \alpha_i([y_1, y_2, \ldots, y_n]\zeta_i).$$

(3) *$\mathfrak{Y}$ is the expression obtained from $\mathfrak{X}$ by substituting for each x_i wherever it occurs the corresponding ζ_i; then*

$$\vdash (t_1, t_2, \ldots, t_p) F_n \beta_1 \beta_2 \ldots \beta_n \delta([y_1, y_2, \ldots, y_n]\mathfrak{Y}).$$

This theorem represents the general result that if we have a function of m variables on $\alpha_1, \alpha_2, \ldots, \alpha_m$ to δ and if for each x_i we substitute a function of n variables on $\beta_1, \beta_2, \ldots, \beta_n$ to α_i, then the result is a function on $\beta_1, \beta_2, \ldots, \beta_n$ to δ. The result is valid when everything depends on certain parameters $t_1, t_2, \ldots, t_p$. It may be generalized by replacing hypothesis (2) by the following: (2′) *For $i = 1, 2, 3, \ldots, m$, ζ_i is a combination of constants, the $t_1, t_2, \ldots, t_p$, and a certain selection $z_1^i, z_2^i, \ldots, z_{p_i}^i$ of the variables $y_1, y_2, \ldots, y_n$ and $\beta_1, \beta_2, \ldots, \beta_n$ are combinations of constants and the $t_1, t_2, \ldots, t_p$ such that, for $\gamma_k^i = \beta_j$ whenever $z_k^i = y_j$, we have*

$$\vdash (t_1, t_2, \ldots, t_p) F_{p_i} \gamma_1^i \gamma_2^i \ldots \gamma_{p_i}^i \alpha_i([z_1^i, z_2^i, \ldots, z_{p_i}^i]\zeta_i).$$

This takes care of the case where the ζ_i do not involve all the $y_1, y_2, \ldots, y_n$.

3.3. (Special consequences of Axioms F to (FK), inclusive.) *The following formulas hold:*

$$\vdash (x) Fxx I \tag{1}$$

$$\vdash (x, y, u_1, u_2, \ldots, u_n) F(Fxy)(F(F_n u_1 u_2 \ldots u_n x)(F_n u_1 u_2 \ldots u_n y)) B_n. \tag{2}$$

3.4. (General consequence of Axioms $(FP)_1$, $(FP)_2$, F.) *If X, $\alpha_1, \alpha_2, \ldots, \alpha_n$, $\alpha_1', \alpha_2', \ldots, \alpha_n'$, β, β' are entities such that*

$$\vdash (x) \alpha' x \supset \alpha_i x \qquad (i = 1, 2, \ldots, n),$$

and

$$\vdash (x) \beta x \supset \beta' x;$$

then

$$\vdash (x)(F_n \alpha_1 \alpha_2 \ldots \alpha_n \beta x \supset F_n \alpha_1' \alpha_2' \ldots \alpha_n' \beta' x).$$

3.5. (General consequence of Axiom (FII), involves also Axioms F–(FK), inclusive.) *If we make the definition*

$$\Xi_n \equiv [x_1, x_2, \ldots, x_n, y](t_1)(x_1 t_1 \supset (t_2)(x_2 t_2 \supset \ldots \supset (t_n)(x_n t_n \supset y t_1 t_2 \ldots t_n) \ldots)), \tag{1}$$

then

$$\vdash (x_1, x_2, \ldots, x_n) F(F_n x_1 x_2 \ldots x_n \Pr) \Pr(\Xi_n x_1 x_2 \ldots x_n), \tag{2}$$

where Pr *stands for "proposition,"* (which, for the purposes of this paper

is to be interpreted as that which is either true or false). By §3.3 (2) this may be generalized to give

$$\vdash (x_1, \ldots, x_m, y_1, \ldots, y_n)F(F_{m+n}x_1 \ldots x_m y_1 \ldots y_n \mathrm{Pr})(F_m x_1 \ldots x_m \mathrm{Pr})(B_m(\Xi_n y_1 \ldots y_n)). \tag{3}$$

Hence if we define, for any constants $X_1, X_2, \ldots, X_n$ and any combination $\mathfrak{X}$ of constants and the variables $x_1, x_2, \ldots, x_n$,

$$(x_1^{X_1}, x_2^{X_2}, \ldots, x_n^{X_n})\mathfrak{X} \equiv \Xi_n X_1 X_2 \ldots X_n([x_1, x_2, \ldots, x_n]\mathfrak{X}), \tag{4}$$

we have the following result: *If* $\mathfrak{Y}$ *is such that*

$$\vdash F_{m+n}X_1X_2 \ldots X_m Y_1 Y_2 \ldots Y_n \mathrm{Pr}([x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n]\mathfrak{Y},$$

then

$$\vdash F_m X_1 X_2 \ldots X_m \mathrm{Pr}([x_1, x_2, \ldots, x_m](y_1^{Y_1}, y_2^{Y_2}, \ldots, y_n^{Y_n})\mathfrak{Y}). \tag{5}$$

The interesting case here is where all the X_i, Y_j are the same.

4. *Defined Functionality.*—It is evident that Axiom F expresses the characteristic property of F. Suppose we define

$$F' \equiv [x, y, z](u)(xu \supset y(zu)).$$

Then F' is a kind of functionality notion, analogous to F. If we take it to be the same as F, then the Axioms F to $(FP)_2$, inclusive (with their consequences, of course), become provable formulas, provided we assume

Axiom (PB). $\vdash (x, y, z)((x \supset y) \supset ((z \supset x) \supset (z \supset y)))$.
Axiom (PC). $\vdash (x, y, z)((x \supset (y \supset z)) \supset (y \supset (x \supset z)))$.
Axiom (PW). $\vdash (x, y)((x \supset (x \supset y)) \supset (x \supset y))$.
Axiom (PK). $\vdash (x, y)(x \supset (y \supset x))$.

(These axioms are intimately related to the corresponding ones for F'.)

As for Axiom $(F\Pi)$, it can probably not be proved for F'. If it be assumed, we shall, with reasonable assumptions as to negation, etc., get a contradiction. This contradiction is considered in Appendix B of Russell's Principles of Mathematics (Cambridge, 1903). It gives rise to the construction of a function f, such that $\vdash F'\mathrm{Pr}\mathrm{Pr}f$ while $P_1^*\mathrm{Pr}f$ is self-contradictory. On the other hand the formula $\vdash F\mathrm{Pr}\mathrm{Pr}f$ is probably not provable, so that we may assume Axiom $(F\Pi)$, as stated, for an abstract F.

Thus if we hold to the suggested interpretations F is a more restricted concept than F', and the assertion $\vdash FXYf$ says more about f than that fx belongs to Y for every x in X. For many purposes it is desirable to retain this more restricted F, and we obtain increased generality by so doing.

5. *The Russell Paradox.*—This paradox arises as follows: Let N stand for negation and Pr for proposition (which means, in this connection, whatever is either true or false). Let

$$\chi \equiv [f]N(ff) \tag{1}$$

(i.e., χ is the property of being non-self-predicable). Then

$$\vdash \chi\chi = N(\chi\chi). \tag{2}$$

Now all this is perfectly correct. In fact without variables we can define

$$\chi \equiv W(BN), \tag{3}$$

and then derive (2) simply by use of Rules W and B. But we do not have a paradox until we have, in addition to (2),

$$\vdash \mathrm{Pr}(\chi\chi), \tag{4}$$

from which it will follow that $\chi\chi$ is both true and false.

If we attempt to prove (4) we find that it follows from the assumptions

$$\vdash FE\mathrm{Pr}N, \tag{5}$$

$$\vdash (x)(Ex \supset FEEx), \tag{6}$$

where $E \equiv WQ$ is the category of all entities. Of these it is useless to deny (6); for it may be proved for F' and Axiom (FII) is not involved. On the other hand (5) is, intuitively, not even plausible. If we replace it by $\vdash F\mathrm{Pr}\mathrm{Pr}N$ the proof of (4) fails. Of course (5) may be made more plausible by replacing N by an entity $\mathfrak{N}$ defined in terms of N and having an interpretation such as the property of being a false proposition, but we shall not get a contradiction unless we make an illegitimate assumption of type (5) for some one of the entities entering into $\mathfrak{N}$.

Under these circumstances there is, I contend, no paradox. The statement (2) is not in itself paradoxical.

A similar explanation applies to certain paradoxes involving relations.

6. *The Epimenides Paradox.*—This paradox is concerned with the application of logic, but deserves consideration none the less, since a logic which cannot be applied is useless. Suppose then we have an intuitive property ϕ and define

$$A \equiv (x)(\phi x \supset Nx).$$

The following are to be regarded as established by experiment:

$$\vdash \phi A \tag{1}$$

$$\vdash (x)(\phi x . x . \supset . x = A) \tag{2}$$

(i.e., all propositions having property ϕ except possibly A are false). This is a possible situation, although it may never have occurred; the question is what can logic say about it.

It is alleged that from (1) and (2) we have

$$\vdash A \supset NA \text{ and } \vdash NA \supset A.$$

Suppose we grant this for the sake of argument. To get a real contradiction we must also have

$$\vdash \mathrm{Pr}A \tag{3}$$

which would follow from

$$\vdash F\mathrm{Pr}\mathrm{Pr}\phi. \qquad (4)$$

But (4), and for that matter also (3), is preposterous.

Other forms of the Epimenides paradox I shall not attempt to discuss. For Russell's propositional paradox see paragraph 4.

7. *Concluding Remarks.*—1. Since the preceding theory does not involve, except in illustrations, any special assumptions as to what the categories of logic shall be, it is not tied to any particular logical theory, new or old, but can serve as foundation for a great variety of them. Thus it is compatible with the engere Funktionenkalkül (Hilbert and Bernays, Grundlagen der Mathematik I, paragraph 4, Berlin (1934)), with the Principia Mathematica (when suitably modified), with Zermelo's Theory of abstract sets (*Mathematische Annalen*, **65**, 261–281 (1908)), or with Heyting's quasi-intuitionistic theories (Berliner Sitzungs berichte, 1930). In fact it would seem to be fundamental to any theory in which distinction of categories is made at all. I am in general agreement with those who believe that any satisfactory logical theory must make such distinction.

2. In relation to Zermelo's theory we may redefine definiteness thus: f shall be definite over M when and only when $\vdash FM\mathrm{Pr}f$. The objections to the idea of definiteness will presumably not then apply. This should be compared with Fraenkel's definition (see, e.g., *Math. Zeit.*, **22**, 254 (1925) —Fraenkel's formulation is equivalent to a definition of definiteness, although he does not phrase it that way), which however rests on a special *ad hoc* definition of function, instead of on a general one, as here.

3. It has been objected that there is no assurance that this theory does not lead to other contradictions. True. However, the theory is no different from any other logical theory in that respect; indeed it is questionable if we shall ever have such assurance, other than that derived from empirical considerations, for any logical theory sufficiently powerful to be of any use. There is, therefore, little to be gained by adopting a Fabian policy in regard to these contradictions. Our friends the physicists have ceased to search for a theory of which they could be sure beforehand that it would explain the universe. We shall do better if we likewise make bold hypotheses which can be modified later, as further research shows their inadequacy. At several points I have made such hypotheses for the sake of simplicity, and propose this theory for examination and, possibly, later modification.

* The results of this investigation were presented in part to the American Mathematical Society on September 9, 1930. Research on this topic was continued while the author was National Research Council Fellow at the University of Chicago in 1931–32. Minor revisions have been made since.

THE CALCULATION OF RELATIVE MULTIPLET STRENGTHS IN A TRANSITION ARRAY

BY GEORGE H. SHORTLEY*

MASSACHUSETTS INSTITUTE OF TECHNOLOGY AND HARVARD UNIVERSITY

Communicated October 13, 1934

In this note we shall sketch the method of calculating relative (dipole-) multiplet strengths in transition arrays to which general formulas are applicable. The results to be given follow from the theory as outlined, e.g., by Johnson,[1] but it is felt that a workable exposition of the general application of the theory will be useful.

Let α and β represent two configurations between which radiative transitions are allowed, and let γ represent a third configuration in which no electrons are equivalent to any in α or in β.[2] Then if we know the relative strengths of the multiplets in the transition array $\alpha \longrightarrow \beta$, we may obtain immediately those in the array $\gamma + \alpha \longrightarrow \gamma + \beta$, where $\gamma + \alpha$ is the configuration which contains all the electrons of γ and of α. We shall give first a general discussion and then illustrate by a particular example. (See footnote 3 for definitions of the spectroscopic terms.)

Let us characterize a term of $\gamma + \alpha$ by the terms of γ and of α on which it is based, i.e., by the quantum numbers L_γ, S_γ, L_α, S_α, L, S. Similarly, a term of $\gamma + \beta$ is characterized by L'_γ, S'_γ, L_β, S_β, L', S'. The strength of the multiplet connecting these terms we denote by $\boldsymbol{S}(L_\gamma S_\gamma L_\alpha S_\alpha LS;\ L'_\gamma S'_\gamma L_\beta S_\beta L'S')$. This strength vanishes unless

$$L_\gamma = L'_\gamma,\ S_\gamma = S'_\gamma,\ S_\alpha = S_\beta,\ S = S'. \tag{1}$$

Let us call

$$\sum_{L,L'} \boldsymbol{S}(L_\gamma S_\gamma L_\alpha S_\alpha LS;\ L_\gamma S_\gamma L_\beta S_\beta L'S) = \boldsymbol{S}(L_\gamma S_\gamma L_\alpha S_\alpha S;\ L_\gamma S_\gamma L_\beta S_\beta S) \tag{2}$$

the strength of the *supermultiplet* $L_\gamma S_\gamma L_\alpha S_\alpha S \longrightarrow L_\gamma S_\gamma L_\beta S_\beta S$, extending the usual meaning of "supermultiplet" to cases in which α and β contain more than one electron. Then *the relative strengths of the multiplets in a given supermultiplet are the same function of* L_γ, L_α, L; L_γ, L_β, L' *as the relative strengths of the lines of the multiplet* $LS \longrightarrow L'S$ *are of* S, L, J; S, L', J'. The relative strengths of the lines of a multiplet are given by the familiar Kronig formulas.[4]

It now remains merely to give the following formula for the relative total strengths of the different supermultiplets:

$$\boldsymbol{S}(L_\gamma S_\gamma L_\alpha S_\alpha S;\ L_\gamma S_\gamma L_\beta S_\beta S) = \frac{(2S+1)(2L_\gamma+1)}{2S_\alpha+1}\,\boldsymbol{S}(L_\alpha S_\alpha;\ L_\beta S_\beta), \tag{3}$$

where $\boldsymbol{S}(L_\alpha S_\alpha;\ L_\beta S_\beta)$ is the strength of the multiplet $L_\alpha S_\alpha \longrightarrow L_\beta S_\beta$ in the array $\alpha \longrightarrow \beta$.

As an example, let us consider some of the supermultiplets of $d^2sp \longrightarrow d^2p^2$. Here $\alpha = sp$, $\beta = p^2$, $\gamma = d^2$. We shall denote the terms of d^2sp based on $d^{23}P$ and sp^1P by $d^2(^3P)sp(^1P)^3SPD$. In this notation each of the supermultiplets allowed by the selection rules (1) is represented by an arrow in the following scheme:

$$
\begin{array}{llll}
d^2(^1S) & sp(^1P)^1P \longrightarrow p^2(^1S)^1S & 1 & (^1P \longrightarrow {}^1S) \\
 & \qquad\quad \searrow\ p^2(^1D)^1D & 5 & (^1P \longrightarrow {}^1D) \\
 & (^3P)^3P \longrightarrow p^2(^3P)^3P & 9 & (^1P \longrightarrow {}^1P) \\
d^2(^3P) & sp(^1P)^3SPD \longrightarrow p^2(^1S)^3P & 9 & (^3P \longrightarrow {}^3S) \\
 & \qquad\quad \searrow\ p^2(^1D)^3PDF & 45 & (^3P \longrightarrow {}^3D) \\
 & sp(^3P)^1SPD \longrightarrow p^2(^3P)^1SPD & 9 & (^3P \longrightarrow {}^3P) \\
 & {}^3SPD \longrightarrow {}^3SPD & 27 & (^3P \longrightarrow {}^3P) \\
 & {}^5SPD \longrightarrow {}^5SPD & 45 & (^3P \longrightarrow {}^3P)
\end{array}
$$

$d^2(^1D)$, etc. (4)

The relative strengths of the multiplets in each supermultiplet are the same as the relative strengths of the lines in the multiplet shown in parentheses at the right of (4), and the relative total strengths of the different supermultiplets are given by the numbers in italics. These are obtained from (3) and the known relative strengths of the multiplets in the array $sp \longrightarrow p^2$, namely,

$$\boldsymbol{S}(^1P;\ {}^1S):\boldsymbol{S}(^1P;\ {}^1D):\boldsymbol{S}(^3P;\ {}^3P) = 1:5:9.$$

For example, for the last entry of (4),

$$\frac{(2S+1)(2L_\gamma+1)}{2S_\alpha+1}\,\boldsymbol{S}(sp^3P;\ p^{23}P) = \frac{(5)(3)}{3}\,9 = 45.$$

The special case of the above results in which α and β are one-electron configurations is already well known; the value $(2S+1)(2L_\gamma+1)$ for the relative total strengths of the supermultiplets here follows from the sum rule. The case in which γ is a single s electron has recently been considered by Ufford and Miller.[5]

These results are being applied by Mr. Leo Goldberg of Harvard University to a number of transition arrays of astrophysical interest.

* NATIONAL RESEARCH FELLOW.

[1] Johnson, *Proc. Nat. Acad. Sci.*, **19**, 916 (1933).

[2] This is an essential condition which must not be transgressed in the application of

these formulas. The only exception is the case of the two-electron array $nl^2 \longrightarrow nln'l'$, for which the allowed multiplets have the same relative strengths as the corresponding multiplets of $nln''l \longrightarrow nln'l'$.

[3] It will be convenient to summarize here the spectroscopic terminology which we shall employ:

Level—A set of $2J + 1$ states characterized by a given J value and having the same energy in the absence of an external field.

Term—A set of $(2L + 1)(2S + 1)$ states characterized by a given L and S.

Polyad—A set of terms of the same multiplicity based on the same parent term in the case of the addition of one electron to a parent ion.

Line—Radiation resulting from all transitions between two levels.

Multiplet—Radiation resulting from all transitions between two terms.

Supermultiplet—Radiation resulting from all transitions between two polyads. A more general definition of this term is given in the text.

Transition Array—Radiation resulting from all transitions between two configurations.

Strength—Sum of the absolute squares of the components of electric moment joining the sets of states in question. The intensity of a single line is proportional to the product of the number of atoms in each state of its initial level, the fourth power of its frequency and its strength.

[4] See Russell, *Proc. Nat. Acad. Sci.*, **11**, 324 (1925), and White and Eliason, *Phys. Rev.*, **44**, 753 (1933).

[5] Ufford and Miller, *Phys. Rev.*, **46**, 283 (1934).

THE LOCALIZATION OF PURE TONES

BY S. S. STEVENS AND E. B. NEWMAN*

HARVARD PSYCHOLOGICAL LABORATORY

Communicated September 25, 1934

A person can localize a tone in space primarily because the sound differs at the two ears in intensity, or in phase, or in both. Experimentation[1] shows that, when a tone is led separately to the two ears, the perceived position of the source can be altered at will by varying either the relative intensity or the relative phase. Localization departs from the median plane (the plane dividing the head and perpendicular to the aural axis) toward the side where intensity is greater and/or where phase leads.

If we limit our consideration to the localization of an actual source of tone in the horizontal plane, then the bilateral symmetry of the observer suggests that similar results ought to be obtained on either side; and experiments[2] show that sounds on one side of the observer are practically never mistaken for sounds on the other. Moreover, the placement of the ears is such that, except for the effects of slight irregularities of the head, there are no cues by which the observer can distinguish tones in front from those behind. Hence, it is justifiable, in considering the problem of rela-

tive accuracy with which an observer can localize tones of different frequencies, to deal only with localization on one side and not to count front-back reversals of localization as errors. This procedure was followed in the present study.

It has been shown that phase-difference is an effective basis for localization at low tonal frequencies.[3] However, phase-difference becomes equivocal at high frequencies, because there is then more than one position of the sound-source which would give rise to the same binaural difference of phase. The presumption is that localization is built up in experience, and that only univocal bases for localization can become established as effective cues.

Phase-difference provides a univocal basis for the localization of all those tones whose wave-length is so long that the difference in phase does not exceed 360° when the sound-source is shifted from a position directly in front to one directly at the side. On the assumption that the head is a sphere of radius 8.75 cm. the critical frequency can be calculated. To a first approximation the difference in the distance which a sound must travel to reach the two ears is

$$d = 8.75(\sin \theta + \theta)$$

where θ is the angle between the azimuth of the sound and the median plane of the observer. Since, neglecting front-back reversals, we are concerned with only one quadrant, localization ought not to become equivocal until there is more than one position within a single quadrant which would give rise to the same phase-difference at the two ears. In this sense equivocality begins at the frequency whose wave-length is equal to the maximum value of d, or 1520 cycles. As the frequency of a tone increases beyond 1520 cycles, cues for localization become less adequate, because a given phase-difference is associated with more and more positions of the sound-source.

Differences in intensity at the two ears result from the sound-shadow cast by the head. Since this shadow is sharper at high than at low frequencies, greater differences of intensity are available as cues for localization at the high frequencies. Measurements of the difference in intensity at the two ears of tones coming from a source at the side of the observer[4] show that the difference is relatively small for frequencies up to 3000 cycles and that above 3000 cycles it rises rapidly to a maximum of 30 db. Therefore, on the basis of difference in intensity alone, we should expect good localization at high and poor localization at low frequencies.

Since the observer can use either the phase-effect or the intensity-effect in localizing tones, he should be successful in localizing both high and low tones. However, throughout a band of intermediate frequencies, both

phase and intensity provide less adequate differences, and in this intermediate region localization should be poorest.

In order to test this deduction we performed an experiment on a platform elevated 9 feet above the roof of the Biological Laboratories of Harvard University. Tones were generated by a loud speaker attached to the end of a 12 ft. arm which was free to rotate in a horizontal plane around the observer. The absence of vertical surfaces at the level of the observer eliminated the possibility of errors due to reflected sound. The tones were presented at each of 13 positions, 15° apart, on the right side of the observer.

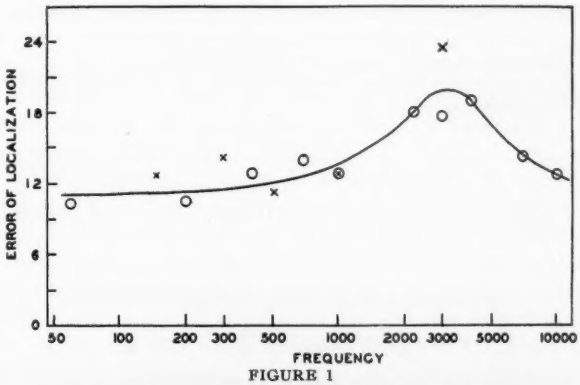

FIGURE 1

Dependence of localization on frequency. The ordinate represents the average of the errors in degrees made by both observers. The crosses are for the shorter series of judgments made in 1933. The circles represent the results obtained in 1934. Note the region of poor localization around 3000 cycles.

The results of the experiment are shown graphically in figure 1. Each circle represents the average of the errors made by both of the two observers. Each observer made 10 observations with the tone at each of the 13 positions at every frequency. The crosses represent some preliminary observations. In computing the errors of localization we took the difference between the reported position of the sound and either the actual position or the corresponding position in the other quadrant. Thus, if the tone was at 30° from the median front position, the localization was regarded as correct if the observer reported either 30° or 150°.

It is clear from figure 1 that the accuracy of localization of all tones below 1000 cycles is approximately the same as that of tones above 7000 cycles, whereas the localization of tones between 2000 and 4000 cycles is relatively poor. This finding is precisely what is anticipated by our previous con-

siderations; although it contradicts the previous general belief[5] regarding the impossibility of localizing high tones.

These results point definitely to a dual mechanism for the angular localization of tones in the horizontal plane. Phase-differences are the primary basis for the localization of low tones, whereas high tones are localized principally by reason of differences in intensity at the two ears. In the intermediate region around 3000 cycles, where the differences in intensity are relatively small and where phase-differences are equivocal, the observer localizes tones with the least accuracy.

* NATIONAL RESEARCH COUNCIL FELLOWS in the Biological Sciences.

[1] O. C. Trimble, *Psychol. Rev.*, **35**, 515–523 (1928).

[2] A. H. Pierce, *Studies in Auditory and Visual Space Perception*, New York, 1901, p. 52.

[3] H. M. Halverson, *Amer. J. Psychol.*, **38**, 97–106 (1927).

[4] J. C. Steinberg and W. B. Snow, *Bell System Tech. J.*, **13**, 247–260 (1934).

[5] H. M. Halverson, op. cit.

ACETYL CHOLINE AND CHROMATOPHORES

BY G. H. PARKER

BIOLOGICAL LABORATORIES, HARVARD UNIVERSITY

Communicated October 8, 1934

A few years ago I published a brief note on the effects of acetyl choline on chromatophores (Parker, 1931). The results detailed in this note, considering the remarkable properties of this material, were in no way startling and were obtained by using only a relatively concentrated solution of the substance (1 part in 1000). After injecting this concentration into Fundulus the fishes darkened slightly, that is, their melanophore pigment became dispersed and in most cases they died. It was clear that this concentration was near the lethal limits for Fundulus. Since then I have learned of the work of Loewi and Navratil (1926), Engelhart and Loewi (1930), Matthes (1930) and Gollwitzer-Meier and Bingel (1933) all of whom have pointed out the destructive effect of blood on acetyl choline and the success with which this destruction can be overcome by the use of physostigmine. Since my first publication Wunder (1931) has reported very briefly on what appeared to be an essentially negative action of acetyl choline on the chromatophores of Rhodeus, and Smith and Smith (1934) have noted no positive effects from this substance when injected into Scorpaena. I therefore resolved to test the matter anew and to employ physostigmine as a protective agent.

Physostigmine itself has been shown to have an effect on melanophores

(Smith, 1931) in that fairly strong solutions of this material when injected into Fundulus will call forth a darkening of this fish. The minimum concentration that will bring about this change is 0.013 per cent and for Fundulus the amount necessary to inject is 0.3 cc. I therefore made up a solution of 1 part of physostigmine in 10,000 parts of Ringer's solution, a concentration less than that which, according to Smith, would excite melanophores and yet stronger than would be needed for the protection of acetyl choline (1 part in 100,000 as stated by Gollwitzer-Meier and Bingel, 1933). Two-tenths cc. of this solution was injected into a number of Fundulus, some of which were in the dark phase (melanophore pigment dispersed) and others in the light phase (melanophore pigment concentrated). Half an hour after these injections the two groups of fishes, which had been kept each in its appropriate environment, showed no change from their original conditions; the dark ones were still dark and the light ones remained light. I concluded from this that the physostigmine at this concentration and in the amount used had no influence on the melanophore pigment of Fundulus.

Light and dark fishes that about 20 minutes previously had received a preliminary injection of physostigmine were then each injected with 0.2 cc. of acetyl choline 1 part in 10,000 of Ringer's solution. In a very few minutes the dark fishes had become much lighter as contrasted with the controls but the light fishes had not changed at all. In a second test the steps just described were repeated except that the dilution used was 1 part in 1,000,000. Again the dark fishes lightened though only slightly while the light fishes remained unchanged. Using 1 part in 100,000,000 the lightening, if it occurred at all, was very little. In a number of instances it certainly did not take place. These results were obtained repeatedly and I concluded from them that acetyl choline has a dispersing action on the pigment of the melanophores of Fundulus and that this action is to be observed in concentrations from 1 part in 10,000 to 1 part in 1,000,000 but not with any certainty at a concentration of 1 part in 100,000,000.

It was a noteworthy fact that many of the Fundulus injected with acetyl choline, 1 part in 10,000, lived only one hour or so after the injection and that when they died they were always in the dark condition. It was also interesting to observe that if fishes appropriately prepared in which a fresh, dark, caudal band had been produced by cutting the tail as already described (Wyman, 1924; Parker, 1934) were injected with even the strongest solution of acetyl choline, the band persisted though the fish itself blanched. In my opinion this result indicates that the exciting agent of the band, the dispersing neurohumor, is momentarily so powerful that the acetyl choline is unable, so to speak, to counteract it.

Very similar results to those just recorded were obtained by the use of physostigmine and acetyl choline on the catfish, Ameiurus. Specimens of

Ameiurus after having received preliminary doses each of 0.2 cc. of physostigmine, 1 part in 10,000, were injected with 0.5 cc. of acetyl choline of different strengths. To 1 part of acetyl choline in 10,000 parts of Ringer's solution the fishes blanched quite observably as compared with their controls. To 1 part in 100,000 the blanching was only slight, and to 1 part in 1,000,000 and 1 part in 100,000,000, no blanching at all was observed.

Through the kindness of Dr. P. B. Armstrong I was enabled to test the melanophores of Fundulus with mecholine (acetyl-beta-methylcholin chloride, Merck), a substance closely related to acetyl choline but much more stable. This material was dissolved in Ringer's solution and injected at different concentrations directly into Fundulus which had received no preliminary treatment with physostigmine. One-tenth cc. of a solution of 1 part of mecholine in 1000 parts of Ringer's solution quickly lightened in a spotty way dark Fundulus, but had no immediate effect upon light ones. All specimens injected in this way sooner or later died. At the weaker dilutions of 1 part in 10,000 and 1 part in 100,000 no effect on the melanophores of the fishes was observed. Mecholine apparently acts on these color-cells as acetyl choline does but rather less effectively.

As a result of these tests it may be stated that acetyl choline when protected by physostigmine causes Fundulus and Ameiurus to blanch, that is, to concentrate their melanophore pigment. Mecholine unprotected also induces this concentration of melanin. It is interesting to note that acetyl choline and mecholine excite in melanophores the same type of reaction as the sympathetic nerve-fibres of these color-cells do.

In my former communication on this subject (Parker, 1931) I stated that the slight action that acetyl choline appeared to have on melanophores was to disperse the pigment and thereby darken the fish. In the present investigation the reverse condition was found to be true. To clear up this contradiction I repeated exactly my former tests and I was interested to find that when a relatively strong solution of acetyl choline unprotected with physostigmine was injected into a Fundulus the fish darkened slightly as formerly reported. This change, however, came on relatively slowly and was almost invariably accompanied by the death of the fish. Heavy doses of acetyl choline are known to have a destructive action on the vertebrate heart and it occurred to me that in my earlier tests the injection of the acetyl choline may have so limited the action of the heart that the resulting loss of circulation had induced a dispersion of melanophore pigment and thereby darkened the animals.

When the heart of a normal light-tinted Fundulus is cut out, the animal naturally soon dies and dies in a moderately dark condition. I believe it is an effect of this kind due to the action of acetyl choline on the heart and circulation that I described in my first communication and not, as I then assumed, the direct action of this drug on the melanophores. In this way

the apparent contradiction in my two papers is to be explained. In my opinion the correct conclusion to be drawn from these studies is that acetyl choline, when protected by physostigmine, and the related compound mecholine alone both induce a concentration of pigment in fish melanophores (light phase of the fish) but only in such relatively strong concentrations (1 part in 1000 and 1 part in 10,000) that neither of these substances can be regarded as of normal physiological significance for the activities of the color-cells in the fishes tested.

Summary.—Acetyl choline when protected by physostigmine and mecholine unprotected both induce a slight dispersion of melanophore pigment in fishes (darkening of the skin).

Engelhart, E., and Loewi, O., "Fermentative Azetylcholinspaltung im Blut und ihre Hemmung durch Physostigmin," *Arch. exp. Path. Pharmak.*, **149**, 1–13 (1930).

Gollwitzer-Meier, K., and Bingel, A., "Der Nachweis eines azetylcholinartigen Stoffes in der Haut," *Ibid.*, **173**, 173–179 (1933).

Loewi, O., and Navratil, E., "Ueber humorale Uebertragbarkeit der Herznervenwirkung. XI. Mittheilung. Ueber den Mechanismus der Vaguswirkung von Physostigmin und Ergotamin," *Arch. ges. Physiol.*, **214**, 689–696 (1926).

Matthes, K., "The Action of Blood on Acetyl Choline," *Jour. Physiol.*, **70**, 338–348 (1930).

Parker, G. H., "Effects of Acetyl Choline on Chromatophores," *Proc. Nat. Acad. Sci.*, **17**, 596–597 (1931).

Parker, G. H., "Cellular Transfer of Substances, Especially Neurohumors," *Jour. Exp. Biol.*, **11**, 81–88 (1934).

Smith, D. C., "The Action of Certain Autonomic Drugs upon the Pigmentary Responses of Fundulus," *Jour. Exp. Zoöl.*, **58**, 423–453 (1931).

Smith, D. C., and Smith, M. F., "Observations on the Erythrophores of Scorpaena ustulata," *Biol. Bull.*, **67**, 45–58 (1934).

Wunder, W., "Experimentelle Erzeugung des Hochzeitskleides beim Bitterling (Rhodeus amarus) durch Einspritzung von Hormonen," *Zeit. vergl. Physiol.*, **13**, 696–708 (1931).

Wyman, L. C., "Blood and Nerve as Controlling Agents in the Movement of Melanophores," *Jour. Exp. Zoöl.*, **39**, 73–132 (1924).